The Role of Information Redundancy in Audiovisual Speech Integration

A Senior Honors Thesis

Printed in Partial Fulfillment of the Requirement for graduation with distinction in Speech and Hearing Science in the undergraduate colleges of The Ohio State University.

By

Michelle Hungerford

The Ohio State University
December 2007

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

When most people think about communication, they think of its auditory aspect. Communication is made up of much more; speech perception is dependent on the integration of different senses, namely the auditory and visual systems. An everyday example of this is when someone tries to have a conversation at a noisy restaurant; a person may unconsciously pay attention to the speaker's facial movements in order to gain some visual information in an imperfect auditory situation. In general, listeners are able to use visual cues in impoverished auditory situations (like a restaurant, or a hearing loss.) However, this process also occurs when the auditory signal provides sufficient information alone.

In 1998, Grant and Seitz conducted a study that found that listeners differ greatly in their perceptions of auditory-visual speech. This study generated a lot of questions about how integration occurs, namely what promotes "optimal integration." Research shows that many factors may be involved: it may be characteristics of the listener, the talker, or of the acoustic signal that influence the amount of integration.

The present study looked at the characteristics of the acoustics, namely whether removal of fine spectral information from the speech signal would elicit more use of visual cues, and thus elicit greater audiovisual integration. The auditory stimuli were degraded by removing the spectral fine structure and replacing it with noise, but retaining the envelope structure. These stimuli were then output through 2-,4-,6-, and 8-channel bandpass filters. Ten listeners with normal hearing were tested under auditory-only, visual-only, and auditory-plus-visual presentations. Results showed substantial auditory-visual integration over all conditions. Also, significant cross-talker effects were found in

the 2- and 4-channel auditory-only condition. However, the degree of integration produced by the talkers was not related to auditory intelligibility. The results of this study have implications for our understanding of the auditory-visual integration process.

Acknowledgments

I would like to thank my advisor, Dr. Janet M. Weisenberger, for giving me the opportunity to work with her on this thesis. I am grateful for the experience in working so closely with her and to receiving her guidance, support, and passion for research. I would also like to thank Natalie Feleppelle for all of her time, support, and patience throughout this experience. I would also like to show my gratitude to my subjects for all of their time spent in the laboratory. Finally, I would like to thank my parents for all of their love and encouragement.

This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

Speech perception is an essential component of language, and in recent years it has been shown to be more complex than originally thought. Speech perception involves sensory systems other than the auditory system, namely the visual system. Although people rely heavily on their auditory systems, the visual system can supplement auditory signals in environments where auditory stimuli alone are not sufficient for speech recognition. Auditory and visual information are added together to form our concept of speech perception. Recent studies have shown that the brain is unable to ignore the visual input, even when it is different from the auditory signal.

One such compelling study done by McGurk and McDonald (1976) showed that perception of the auditory signal is dependent on information from the visual system. They discovered what is called the McGurk effect; it occurs when different visual and auditory speech stimuli are presented at the same time. The goal of the study was to evaluate how the listener integrated the differing stimuli, and whether one sensory modality would dominate the other. In McGurk and McDonald's study, the subjects received both auditory only and auditory plus visual situations in which to perceive information. The findings of the study concluded that the different stimuli are either fused or combined together to form a new response, which is neither the same as the auditory nor visual stimulus, but rather a mixture of both. For example, if a listener hears the auditory signal of the bilabial consonant [ba] and sees the visual stimulus of the velar consonant [ga], the listener perceives the alveolar [da], which is a fusion of the auditory and visual stimuli because its place of articulation lies between bilabial and velar. The

reverse also occurs; if the auditory stimulus is [ga] and the listener sees the visual stimulus [ba], the listener will interpret the sound as [bga]. This difference in integration is due to the fact that the visual bilabial information is much more noticeable because of its highly visible frontal placement, as compared to a velar placement at the back of the oral cavity, and due to this the visual and auditory stimuli lack fusion. The results of McGurk and McDonald's study demonstrate that cues from sensory systems other than the auditory system are factored in to speech perception and that these cues are unable to be ignored.

This study left many unanswered questions about speech perception. The biggest mystery left is the specifics of the integration itself. The main question that has dominated the future research of integration is what promotes optimal integration; does clear highly intelligible speech assist in the integration process, or is it ambiguity in the speech that creates better integration?

Auditory Cues for Speech Perception

Studies by Shannon and his colleagues (1995) provided evidence that the auditory signal is very "redundant", or contains much more information than is necessary in order to identify the speech sound. An acoustic speech signal has cues for place, manner and voicing in both temporal and spectral aspects of the waveform. For manner, these cues include relative intensity of formants and formant frequency changes. The main cue for place of articulation is formant transitions. For voicing, they are duration of the sound and voice onset time. In order to reduce redundancy, Shannon removed some spectral information while holding the temporal information from speech stimuli constant. He

achieved this by replacing selected spectral information with band-limited noise (Shannon et al., 1995). Shannon discovered that subjects were able to recognize speech with substantial accuracy with only three noise bands modulated by the temporal envelope. Speech recognition performance improved steadily as the number of noise bands was increased. With this study, Shannon demonstrated that auditory signals are very redundant and that speech sounds can be identified even when a large amount of spectral information is removed.

Shannon went on to study the effects on the temporal cues of specific forms of spectral degradation, in an attempt to better define the parameters of speech recognition for consonants, vowels, and sentences (Shannon, Zeng, & Wygonski, 1998). This study contained four experiments that dealt with the spacing of cutoff frequencies, warping the spectral distribution of envelope cues, frequency shifting envelope cues, and spectral smearing. Overall, these experiments showed that, for four bands, the frequency alignment of the analysis and carrier bands is critical for good performance (Shannon et al., 1998). Experiments I and IV showed that an overlap in carrier bands was not critical for speech recognition; only when the bands were broadly overlapping did spectral smearing occur causing a decrease in speech recognition. Experiments II and III, however, demonstrated that when spectral cues are warped speech becomes completely unintelligible. Shannon's study also showed that recognition of vowels is affected more than that of consonants when spectral warping was employed.

Remez et al. (1981) studied the role of redundancy within a speech signal by degrading the signals into sine-wave speech. The speech signal was degraded into three time-varying sine waves that represented natural speech by following the formant

structure of the signals. However, the sinusoids differ from natural speech because of the formant structure. Remez's study speculated that the sine-wave speech should be detected by the listener as three separate tones as opposed to human speech. To test this, three separate conditions were used in which the subjects were given differing levels of information about the stimuli. Only the third group was given extensive information about the stimulus, including the actual wording. The second group was told that they would hear a computerized sentence. The first group was only asked to give impressions of the stimulus and was told nothing else. The study showed that primed listeners were able to pay attention to, and identify the reduced-cue speech. Because sine-wave speech can be perceived as speech, Remez concluded that speech cues previously thought to be the basis of speech perception may only be secondary structures in the process.

Visual Cues for Speech Perception

The previous experiments focused on the auditory signal and its contribution to speech perception regarding features such as place, manner and voicing. However, other studies have attempted to determine the impact of visual cues and their role in speech perception. Other than lip, jaw and tongue movement, visual cues may consist of the talker's eye movements as well as movement of the head (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004). Unlike auditory cues, visual cues are limited in the types of information they provide to the speech perception process; the only substantial information they carry is for the place of articulation.

The problem with visual cues, as stated above, is that they provide no information about consonant voicing, and only limited information about consonant manner. Many

sounds are produced at the same place in the mouth, such as the bilabials /p/, /b/, and /m/, or the velars /k/ and /g/. This poses a problem, because without auditory cues to accompany the visual ones, some sounds are completely indistinguishable from one another. Sounds like /k/ and /g/ are not only indistinguishable from one another visually, but they can also be very difficult to distinguish from other sounds made toward the back of the mouth, such as alveolar or glottal sounds, by using visual cues alone. In order to study visual cues, researchers have grouped together similar sounds by the visual movement required to make the sound; these groups are called visemes (Fisher, 1968). Visemes allow for a person to distinguish among groups of sounds, but not the individual sounds within a group (Jackson, 1988).

A great deal of research has been done in regards to visual aspects of vowels and consonants. As mentioned above, the greatest information provided by visual cues is the place of articulation, which is valid for consonants. The most commonly grouped visemes are /p,b,m/, /f,v/, and /θ/ because the movements of the mouth are very general for these groupings (Jackson, 1988). For vowels, mouth shape is the most definitive quality with regards to place of articulation for categorization of visemes. Of the mouth shape features, the most prominent for speechreading are lip extension versus rounding, and vertical lip separation (Binnie, Jackson, Montgomery, 1976). Productions of both consonants and vowels give visual cues that aid in speechreading.

The main problem with visual cues, however, is that they are highly dependent on the talker, which can make speechreading a daunting task. The formation of visemes is related to the characteristics of the talker and the ease with which a talker can be speechread (Kricos and Lesner, 1982). Kricos and Lesner also found that talkers who are

easier to speechread often display a wider variety of viseme categories. Talkers who are easier to speechread display more vowel viseme categories as well (Kricos & Lesner, 1982). In general, talkers who are easily speechread are those who display the most viseme categories.

Auditory-Visual Integration Theories

In the following discussion, "audio-visual integration" refers to the process undergone by the listener during which he or she takes both the auditory and visual stimuli and combines them together (Grant, 2002). Several researchers have developed theories to explain the process of auditory-visual integration. Grant (2002) discussed the validity of different models of speech perception in predicting audiovisual integration performance; the two models he compared are the Prelabeling Model of Integration and the Fuzzy Logic Model of Perception. The Fuzzy Logical Model of Perception (FLMP) proposed by Massaro (as cited in Grant, 2002) was constructed in order to account for discrepancies in the predicted audiovisual score and the obtained scores for auditory and visual situations. The basis of this theory is that stimuli arriving in different sensory channels are each processed prior to being combined and integrated. This creates a summary description for the information. The summary descriptions are taken and compared to descriptions within the memory in order to determine how well the cues align with the information stored in the memory. This model states that after the brain judges the information, then the stimuli are all integrated together, forming perceptual alternatives. However, one problem with this model is that, according to Grant's findings, it is as likely to overestimate the integration rate as underestimate it. In his

paper, Grant states that the Fuzzy Logic model is less reliable because it underestimates the listeners' true integration abilities, whereas the Prelabeling model predicted more accurately.

The Prelabeling Model of Integration (PRE), formulated by Braida (as cited in Grant, 2002) postulates that integration of auditory and visual information occurs very early in the process. This model uses the auditory-alone and visual-alone performance to predict what audiovisual performance will be. The basis of Braida's theory is that response levels for auditory-visual information should be better than those of just auditory or just visual cues. The PRE model defines integration by the amount of audiovisual integration that the listener is able to produce; the higher the listener's audiovisual scores, the more efficiently the listener is integrating. In contrast, listeners with audiovisual scores lower than predicted by the model are considered poor integrators. For Grant's study, the PRE model showed that, as far as integration is concerned, some subjects were just more efficient than others. Grant states that this model seems to be better equipped to predict integration because, when tested on hearing-impaired individuals, it accounted for individual differences within their speech perceptions.

These were the two models used by Grant and Seitz (1998) in which they looked at how audio and visual information are integrated by listeners and how the levels of integration among listeners varies. This study took into account hearing loss, visual acuity, vocabulary, and language competence in order to explain differences across subjects for auditory-visual integration. This study used several different measures for integration, such as correct identification of sentence stimuli or nonsense stimuli, as well

as making those stimuli discrepant or congruent. The results of this study again showed that listeners fuse auditory and visual information together during speech perception; this was demonstrated by the fact that audiovisual scores for consonant recognition were higher than the scores for either audio or visual alone. However, this study also showed that integration is not a clean-cut process, but rather that integration efficiency varies across listeners. Even though audiovisual scores were higher in general, the amount of integration efficiency was very different across listeners.

The Role of Auditory Information in Audiovisual Speech Integration

These studies, although they have answered many questions about how speech perception works, have left a few unanswered questions as well. One unanswered question is exactly what happens to audiovisual integration as the quality of incoming stimuli is changed. One way to study this process is to examine reactions to a closed set of stimuli under different conditions and levels of degrading. In such an approach it is easy to track changes in the stimulus as the auditory signal becomes more and more degraded.

The present study examines the ways in which people integrate auditory and visual stimuli when the auditory stimulus is altered. By systematically removing greater amounts of auditory information, we can see how response patterns are altered when visual stimuli are added. The auditory stimuli were degraded in a similar manner to Shannon et al. (1995); the auditory stimuli were degraded into two spectral bands, four spectral bands, six spectral bands, and eight spectral bands of speech signal effectively reducing the speech signal information. The listeners were tested under all three of the

following conditions: 1) degraded auditory only, 2) visual only, and 3) degraded auditory + visual. Analyses of specific identification responses in confusion matrices were performed in order to determine the exact information that best promotes audiovisual integration. Results should provide insights into the integration process and will help clinicians form better aural rehabilitation programs.

Chapter 2: Method

Participants

There were 15 total participants in the study. Of the 15, five were talkers; there were three female and two male talkers who ranged from 20 to 23 years old. All the talkers reported normal hearing as well as normal or corrected vision. They produced the stimulus set of eight monosyllabic words in front of a video camera. The other ten participants were observers. Of the observers, two were male and eight were female, ranging from 20 to 23 years of age. Five of the observers were Speech and Hearing Science majors doing undergraduate work, and the others had other majors of undergraduate study. Four of the Speech and Hearing majors had some knowledge of the McGurk effect. All observers were tested for normal hearing and all reported normal or corrected vision. Three participants received academic credit for participation in the study, six received $80 compensation for participation, and the rest participated voluntarily.

Interfaces for Stimuli Presentation

The auditory stimuli were presented through TDH 39 headphones, and the visual stimuli were presented from a DVD player on a 20 inch video monitor.

Stimuli Selection

The stimuli set for this experiment consisted of eight CVC syllables. These stimuli satisfied the following conditions:

1. All stimuli included the same vowel, one that does not include lip rounding, /ae/.
2. All stimuli were known to elicit McGurk responses.
3. All stimuli were presented citation-style (the stimulus alone).
4. The stimulus pairs differed only in the initial consonant (minimal pairs).

Stimuli:

The stimulus set described above was observed by the listeners under each condition. This set includes both single and dual-syllable stimuli. For the dual-syllable stimuli, the first stimulus of the pair is the auditory stimulus, and the second is the visual stimulus. These were presented simultaneously in order to elicit the McGurk effect.

Single-syllable stimuli:

- Bilabial: mat, pat, bat
- Alveolar: zat, sat, tat
- Velar: gat, cat

Dual-syllable stimuli:

- cat-pat
- pat-cat
- bat-gat
- gat-bat

Stimulus Presentation

Audio Signal Degrading:

Auditory stimuli in this experiment were degraded in a manner similar to that created by Shannon et al. (1998). The talkers produced a set of eight monosyllabic stimuli a total of five times each. The stimuli were recorded through a microphone and fed directly into a computer, which allowed for the files to be stored in .wav format. The files were put into a subroutine created by Bertrand Delgutte in MATLAB 5.3. The subroutine begins with two stimuli: the input speech waveform and a broadband noise. The program exchanges the amplitude envelope waveform and fine structure of the two stimuli. It then keeps the waveform containing noise fine structure and speech envelope characteristics. The waveforms were then filtered into two bands, four bands, six bands, or eight spectral bands where the bandwidths are chosen to provide equal spacing in basilar membrane distance. The cutoff frequencies for the two spectral bands were: 80 Hz, 1,877 Hz, and 19.2kHz. The cutoff frequencies for the four spectral bands were: 80 Hz, 518 Hz, 1,877 Hz, 6,097 Hz, and 19.2 kHz. The cutoff frequencies for the six spectral bands were: 80 Hz, 315 Hz, 814 Hz, 1,877 Hz, 4,139 Hz, 8,953 Hz, and 19.2 kHz. The cutoff frequencies for the eight spectral bands were: 80 Hz, 238 Hz, 518 Hz, 1,010 Hz, 1,877 Hz, 3,404 Hz, 6,097 Hz, 10,840 Hz and 19.2 kHz. Auditory syllables were thus reduced to a broadband noise fine structure that is modulated by the temporal envelope of the original recording of the speech stimulus.

Digital Video Editing:

The visual stimuli used in this experiment were created using two male and three female talkers recorded by a digital video camera. The talkers repeated a list of eight stimulus words a total of five times each. The visual and auditory stimuli were then

downloaded to the software program Video Explosion Deluxe, which allowed for the clips to be edited. This program allows for any auditory clip to be dubbed onto any video clip, which allowed for the creation of incongruent stimuli, with different auditory and visual components. This allowed for the presentation of McGurk-type integration stimuli. The present study paired visual and auditory stimuli from the same talker. Randomized lists were created for all four conditions in order to reduce the possibility of effects that can occur from order of stimulus presentation. A set of videos was created from these lists, each with sixty stimulus clips. The program Sonic MY DVD was used to burn the individual videos created in Video Explosion Deluxe to DVDs. Three randomized lists were created for each talker for each of the four conditions, which resulted in the creation of sixty DVDs.

Procedure

Testing Setup:

Testing was conducted in The Ohio State University Speech and Hearing Department. The lab provided a quiet environment for testing. Within the lab was a sound-attenuated booth, with a chair placed against the back wall. The participants faced a glass window in the booth through which they could watch for the visual stimuli on the television monitor placed outside. While seated, the participants were roughly four feet from the television monitor. Auditory stimuli were presented through TDH 39 headphones. The examinees' responses were transmitted through an intercom system to the examiner on the outside of the booth.

Eight of the listeners were also tested on the lab computer via headphones for auditory stimuli and the computer monitor for visual stimuli. This was done to determine if the abundance of "hat" responses was a product of additional noise added to the auditory signal by the amplifier for the headphones. Each of the eight participants were retested on an auditory-only or auditory plus visual condition that he/she had a high rate of "hat" responses during initial testing. These extra trials were varied across talkers and number of channels.

Testing Presentation:

Before testing began, each listener was first given a hearing screening to ensure normal hearing thresholds. Each listener was also given a written explanation of the experiment. A verbal explanation was also provided, with a more detailed explanation at the listener's request. These explanations provided the listener with the information that they were to be tested under three conditions (auditory-only, visual-only, and auditory + visual) for each talker for each level of auditory degradation (2-channel, 4-channel, 6-channel, 8-channel). The listeners were told that these conditions would be randomized and included: visual-only, auditory-only, and auditory + visual. The listeners were told that for visual-only they would only be able to watch the television screen and auditory stimuli would not be presented through the headphones. For auditory-only they were only able to listen to the headphones and the monitor was turned off. For auditory + visual, the listeners were able to both watch the television screen and listen through the headphones. The listeners were instructed that each of the stimuli ended in "at" and that only the beginning consonant would change. They were told that any beginning consonant or combination of consonants was a valid response. The listeners were also

instructed that the consonant or consonant clusters did not have to form a known word but could be a nonsense syllable. The listeners were instructed to respond to each of the 60 stimuli on every CD by repeating the syllable that was perceived.

Testing Procedure:

Each listener was tested over approximately 10 hours, in multiple sessions that lasted between one and two hours each. All ten participants were tested in each of the three conditions for each of the five talkers in each of the four levels of degradation, resulting in 60 trials for each condition for each participant. The order of presentations was randomized across both level of degradation as well as the condition for each listener.

Chapter 3: Results and Discussion

The results for two different types of stimuli were analyzed. One type was the congruent (or single-syllable) stimuli, in which the same stimulus was presented to both auditory and visual modalities and percent correct performance was measured across all conditions (degraded auditory-only, visual-only, degraded auditory plus visual). The degree to which the auditory plus visual performance was improved over the auditory-only or visual-only performance served as a measure of integration.

The second type was incongruent (or dual syllable) stimuli, in which the auditory and visual stimuli differ from one another. These responses were not recorded for percent correct because there is no "correct" response for the differing stimuli. Instead, the responses were recorded into three categories: auditory (the response was the same as the auditory stimulus), visual (the response was the same as the visual stimulus), and other (the response differed from both the auditory and visual stimuli.)

Percent Correct Performance

Figure 1 shows the results for percent correct identification for visual-only, auditory-only, and auditory plus visual presentations across all four conditions (2-channel, 4-channel, 6-channel, and 8-channel.) The results shown are averaged across talkers and subjects. This figure indicates several things. The first point worth noting is that visual-only performance is consistent across all channels. This is expected, because the varying factor among the channels is only in the auditory information, which should not affect the visual-only condition. Another point worth noticing is that auditory-only

performance systematically increases with the number of channels. This implies that auditory perception increases when more auditory information is available to the listener. From 2-channel to 4-channel, speech recognition improved by 12%. It improved 4% from 4-channel to 6-channel, and 12% from 6-channel to 8-channel.

Like the auditory-only scores, auditory plus visual scores also increased as the number of channels increased. However, the interesting information with regard to auditory plus visual performance is the percent improvement of the auditory plus visual scores over the auditory-only in each condition. In the 2-channel condition, the auditory plus visual has a 10% improvement over the auditory-only condition. In the 4-channel as well as the 6-channel there is a 14% improvement, and there is a 7% improvement in the 8-channel. This increase in performance is relatively similar over all four conditions with a slight decrease in the 8-channel condition. This decrease suggests that the visual stimulus may not provide as much additional information because the auditory signal is already redundant. The results for the other three channels (the 2-, 4-, and 6-channel) show that the addition of the visual stimulus adds new information to the overall signal that is not available in the auditory-only condition.

Figure 2 represents the percent correct scores for the visual-only condition. The scores were averaged across talkers. This graph shows that there is only a small amount of variability among the talkers in the visual-only condition. Figure 3, however, shows great variability across talkers in the 2-channel auditory-only condition. Talkers LG and PV are much more intelligible than the other three talkers. MO is slightly more intelligible than both JK and KS, who are about evenly intelligible. Figure 4 shows the percent of correct responses across talkers in the 4-channel condition. This figure shows

that all talkers have improved intelligibility over the 2-channel condition. Although LG and PV are still yield higher scores than the others, there is much less variability across talkers, and only JK's scores are much lower than the rest. Figure 5 shows even less variability among talkers in the 6-channel condition, and all the scores are relatively close to one another. Figure 6 is in accordance with Figure 5, showing even less variability among talkers in the 8-channel condition. This figure shows high intelligibility across the board.

McGurk Type Integration

Figure 7 shows the responses for the incongruent stimuli. Auditory response scores are the lowest across channels; however, in the 8-channel condition, auditory response rate is much closer to the other response rates. The visual response is the highest response rate in the 2-channel condition, but shows a steady decrease as the number of auditory channels increase. Auditory plus visual rates are fairly steady across all frequencies, with the highest rates in the 4-channel and the 6-channel conditions. As intelligibility increases, so does the rate of auditory responses. Figure 8 analyzes the "other" responses represented in Figure 7 to assess the amount of integration that occurred. The lowest scores achieved were combinations (when the listener combines the beginning consonants of both words for a response). This finding is congruent with previous research on the subject; combination response rate is low due to the fact that these consonant clusters are not part of Standard American English. Fusion responses (where the listener fuses the beginning consonants together to form a completely different response), although much higher than combination responses, were also low across all

channels. This finding is somewhat baffling because previous studies showed fusion integration near 50%-60% of all responses. The present study found 27% fusion response in the 2-channel, 30% in the 4-channel, 35% in the 6-channel, and 29% in the 8-channel. The results of the present study suggest that removing any auditory information may be harmful. However, there is one concern within this study. It is the high percentage of "hat" responses, which were produced by all subjects across all stimuli. "Hat" was not classified as a fusion response because the location of production is glottal, which does not fall in between the bilabial and the velar.

Results obtained from the computer trials are inconclusive regarding the abundance of "hat" responses. All but one subject had less "hat" responses during the computer trial; however, the degree of responses between the booth and the computer varies greatly among subjects. Both AS and MG had only less "hat" response during the computer trial, whereas MT and KB had a drastic decrease in "hat" responses. NO actually gave five extra "hat" responses during the computer trial that he did not give during the booth trial. Also, during auditory plus visual conditions during the computer trials, the "hat" responses were all given for McGurk stimuli. This is an interesting discovery that may imply integration of some kind as opposed to noise interference. Although these results seem to suggest that the amplifier may be causing some interference, further study is needed to determine the degree of interference.

Figure 9 investigates the low levels of fusion response by looking at the response rates for different talkers. This chart shows a substantial difference between talker LG, who has a very high fusion rate compared to the rest, and JK, who has a very low rate. Figure 10 breaks down the fusion responses over all channels by subject. This figure

reveals a very incongruent fusion response pattern across the subjects. Listeners JL and MH have the highest rates compared to subjects SP, KB, and KV, who all have very low levels of fusion response.

Figures 11 and 12 show the amount of integration in the 2- and 4-channel conditions for each talker. In the 2-channel, talkers MO and KS offer the greatest amount of integration. In the 4-channel condition, talkers MO, KS, and JK offer the most integration effects. These results would seem to imply that the worst talkers in the auditory condition offered the most integration in the auditory+visual condition.

Confusion Matrices

The overall results from the present study raise several questions requiring further examination. Confusion matrices were constructed in order to look at the types of errors made by the subjects, and these matrices can be found in the Appendix of this document. One matrix was made for each condition (2-,4-,6-, and 8-channel) across subjects and talkers. In general, the types of responses are fairly stable throughout the four conditions, but as expected, as percent correct scores increase confusions are less prevalent. The 2-channel matrix delivers the most interesting information for the present study.

In the 2-channel condition, participants correctly perceived “bat” 58% of the time. The group most confused for was the other bilabials “pat” and “mat”. “Mat” accounted for 9% of those responses, showing that voicing and place of articulation were preserved. This confusion may be attributed to the high-frequency energy of “bat” being lost in the degraded signal, and confused for the low-frequency energy of the nasal. “Cat” and “gat” made up 5% of the responses, which shows that the high-frequency energy of manner of

articulation was also conserved. Another high frequency response was “that”, which is consistent in voicing. The prevalent rate of “fat” responses is interesting because it shares no place, manner or voicing characteristics with “bat”. Also, its spectrogram looks different; the spectrogram of “fat” contains a broad energy across frequencies, whereas “bat” lacks most of this energy. This response rate may occur because the absence of spectral fine structure replaced by noise caused the signal to sound more like noise, which may have been confused for the spectral energy throughout the voiceless fricative /f/.

“Pat” was correctly identified 55% of the time in the 2-channel condition. “Bat” was confused for “pat” 5% of the time, showing that place and manner of articulation were preserved within the signal. This difference may be due to a loss of voice onset time information. The longer duration of voice onset time for “bat” may be lost in the degrading, causing the listener to perceive the shorter voice onset time in the low frequencies of “pat”. “Cat” made up 11% of the responses, showing preservation of manner and voicing. Like “pat”, “cat” also has a shorter low frequency voice onset time that is similar to that of “pat”. “Fat”, again, had a high response rate, retaining voicing information.

“Mat” was correctly identified 73% of the time. “Bat” was heard 5% of the time, which shows that manner and voicing were preserved. “Mat”, like “bat” has a longer voicing onset time in both low and high frequencies, which may have been evident in the signal. “Nat” was also perceived 5% of the time, showing again that manner and voicing were well preserved. “Nat” has some more high frequency components to its voice onset time than “mat” which may have caused confusion due to the degrading.

"Gat" was correct 36% of the time. It was confused for "bat" 30% of the time, which is the largest incorrect response rate for this stimulus, showing that both manner and voicing were preserved. "Gat" has more high frequency energy available during voice onset than "bat", which again could be a factor of the degrading. "Cat" made up 10% of the responses, showing that place and manner were both retained. The discrepancy between the two is that "cat" has a shorter voice onset time than "gat" in the lower frequencies. "That" was perceived 4% of the time, which is consistent in voicing.

"Cat" was correct 67% of the time. It was confused for "pat" 16% of the time and both manner and voicing are retained in the signal. "Pat" also has a voice onset time similar to "cat", but with less high frequency information available. "Tat" was perceived 4% of the time, which suggests again that manner and voicing are intact in the signal. "Fat" was also perceived 4% of the time, which is consistent with voicing characteristics.

In general, the alveolar stimuli had the lowest correct response rate. "Zat" was correct only 8% of the time, which is the lowest correct response rate. "Zat" was confused for "bat" and "mat" 20% of the time, which shows that voicing cues are present in the signal. Other alveolar sounds were perceived 9% of the time which is consistent with place of articulation; 5% of those alveolar sounds were also voiced. "Fat" and "that" make up 14% and 21% of the responses, which shows that manner cues are very pronounced in the signal, as well as voicing.

"Tat" was correct only 15% of the time. "Pat" and "cat" made up 31% and 32% of the other responses, so manner and voicing were largely preserved. The other large response rate was "fat", which also kept with voicing characteristics of the signal.

"Sat" was correct 18% of the time. The other fricatives make up 66% of the other responses. This shows that a majority of the time, manner of production is preserved. The 54% labiodental "fat" response may be due to the sound signal available under 4kHz in the word "fat". The sound spectrograms of "fat" and "sat" differ from each other because "fat" has more information under 4kHz than "sat". The degrading of the auditory signal may have caused the listener to perceive more energy in the lower frequencies, thus causing the high confusion rate of "fat" for "sat". It was confused for "bat" 11% of the time. This may be due to loss of information in the higher frequencies due to degrading, because "sat" is spectrally different from "bat" in that it has much more high frequency activity.

The glottal "h" response had a high prevalence for all stimulus words, except for "cat". The remainder of the stimulus words had a range of 4-16%. This high response rate for "hat" is troubling, and may be due to the substitution of noise for speech fine structure, which may be perceived as noise.

In general, the highest response rates for incorrect responses retained some combination of place, manner, and voicing cues. The main manner cue that likely allows for confusion is the duration of the noise. Stops have a burst of noise, affricates a sharp onset, and fricatives a long duration of noise. The replacement of spectral fine structure with noise may be causing these cues to be less noticeable within the auditory signal. The place of articulation cue that accounts for confusion is the frequency of the noise. Many of the stops were confused for fricatives, showing a discrepancy in where the listeners perceived the spectral peak. Again, this may be a side effect of the type of degrading used for the study because the noise structure may be inhibiting the listeners

from discerning the spectral information. The discrepancies in voicing are caused by timing issues, including noise duration and voice onset time. These subtle cues may be overridden by the noise structure as well.

Statistical Analysis

Statistical Analysis (ANOVA) revealed many significant findings. In order to calculate significant differences across channels and presentation conditions, a two-factor, within subject, ANOVA was performed. There was a significant main effect of number of channels in the degraded auditory stimulus, $F(3, 144) = 76.53$, $p<.001$, $r^2 = .61$. The Pairwise Comparisons show significant differences among all pairs of channels, except for the 4-channel and 6-channel comparisons. There was also a significant main effect of presentation condition, $F(2, 96) = 638.11$, $p<.001$, $r^2 = .93$. The Pairwise Comparisons show significant differences among all three of the presentation conditions. There was a significant interaction of channels by presentation condition, $F(6, 288) = 27.45$, $p<.001$, $r^2 = .36$.

A set of analyses was conducted to look at differences across talkers in the 2-channel, 4-channel, 6-channel, and 8-channel conditions. In the 2-channel, there was a significant effect of talker, $F(2.9, 26.3) = 33.31$, $p<.001$, $r^2 = .79$. The Pairwise Comparisons show that LG and PV performed differently from the other talkers. In the 2-channel condition, there is also a significant interaction of talker and condition, $F(8, 72) = 11.243$, $p<.001$, $r^2 = .56$. In the 4-channel analysis, there was a significant main effect of talker, $F(4, 36) = 9.47$, $p<.001$, $r^2 = .50$. The follow-up Pairwise Comparison showed that talker JK performed differently from the rest of the talkers. There was also a significant interaction effect of talker and condition, $F(8, 72) = 3.913$, $p<.001$, $r^2 = .303$.

In the 6-channel analysis, there was a significant effect of talker, $F(4, 36) = 4.8$, $p<.003$, $r^2 = .35$. However, in the 6-channel condition, the Pairwise Comparison revealed that none of the talkers performed differently from the others. There was also a significant interaction effect of talker and condition, $F(8, 72) = 2.410$, $p<.023$, $r^2 = .211$. In the 8-channel analysis there were no differences or main effects across talkers.

Chapter 4: Summary and Conclusion

The results of this study show that as the amount of auditory information available to the listener increases, the subjects perform steadily better. Removing auditory information from the stimulus, however, does not seem to greatly affect integration.

However, removing auditory information seems to have a profound effect on perception of stimuli. The lack of auditory information seems to affect the stimuli from some places of articulation more than others, such as the low percent correct rate for alveolar stimuli. Bilabial stimuli had the greatest rate of correct responses, followed by the velar stimuli, and then the alveolar stimuli. Differences in spectral fine structure of some stimuli, such as “sat” and “fat”, are a cause for confusion when fine spectral information is not available. Place, manner and voicing cues are responsible for the confusions, and these cues may be lost due to the noise structure of the speech signals. When confusions take place, however, most are consistent with place, manner, or voicing characteristics of the original signal. Place cues seem to be the least prevalent. Also, talker characteristics are a large factor in integration, but so are differences among listeners, as displayed in Figure 10, which is consistent with previous research done by Grant and Seitz (1998).

Results from this study indicate that removing auditory information in this manner does not necessarily affect the degree of integration because other factors, such as talker and listener differences, have a profound effect on the integration process. Further study is needed in order to determine how the degree of benefit differs across auditory

conditions as well as talkers. Additional studies are needed to address the high prevalence of “hat” responses as well as how the degree of benefit changes across talkers and conditions in order to study the low McGurk effect observed.

References

Binnie, C.A., Jackson, P., & Montgomery, A. (1976). Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders*, 41, 530-539. (cited in Jackson, 1988)

Braida, L.D. (1991). "Crossmodal integration in the identification of consonant segments," *Quarterly Journal of Experimental Psychology*, 43A (3), 647-677. (cited in Grant, 2002)

Fisher, C.G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 12, 796-804. (cited in Jackson, 1988)

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America*, 112 (1), 30-33.

Grant, K.W. & Seitz, P.F. (1998). Measures of the auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, 04, (4), 2438-2449.

Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception:

Visemes and coarticulation. *The Volta Review*, 90 (5), 99-114.

Kricos, P.B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *The Volta Review*, 84, 219-225. (cited in Jackson, 1988)

Massaro, D.W., & Cohen, M.M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception. *The Journal of the Acoustical Society of America*, 108, 784-789. (cited in Grant, 2002)

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.

Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perceptions and Psychophysics*, 66 (4), 574-583.

Remez, R.E., Rubin, P.E., Pisoni, D. B., Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-950.

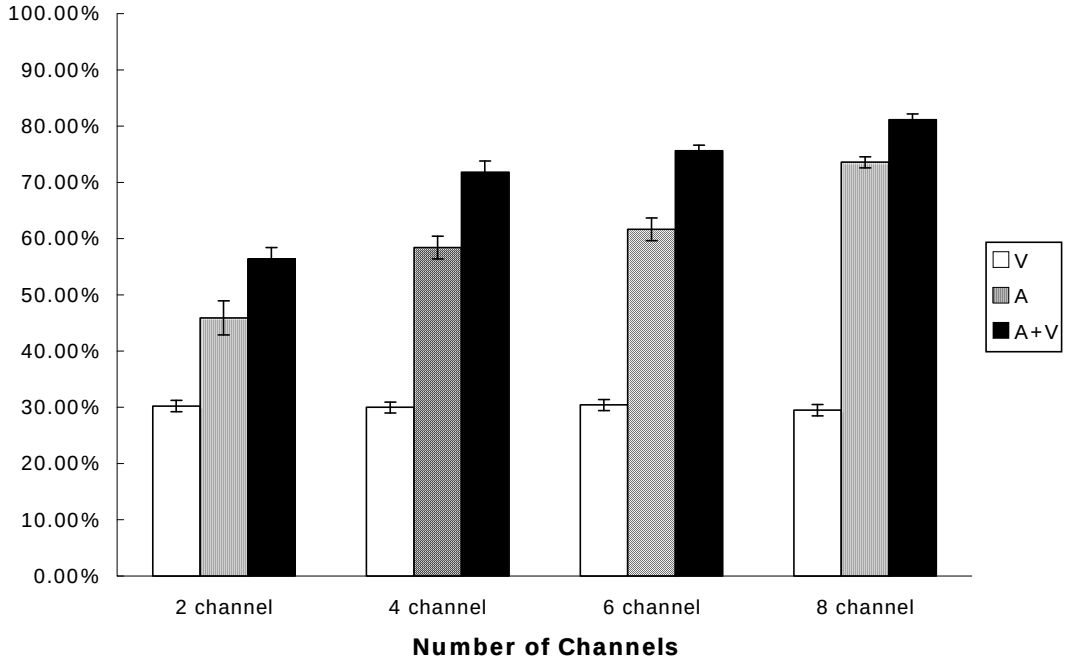
Shannon, R.V., Zeng, F. G, Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily visual temporal cues. *Science*, 270, 303-304.

Shannon, R.V., Zeng, F. G., Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of*

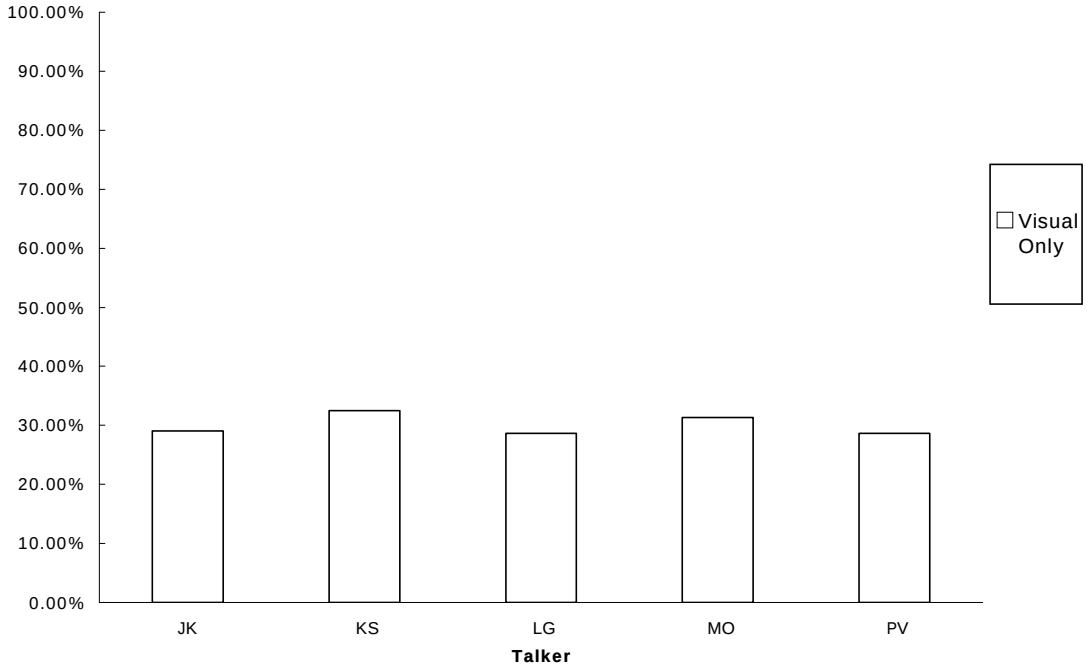
America, 104 (4), 2467-2475.

List of Figures

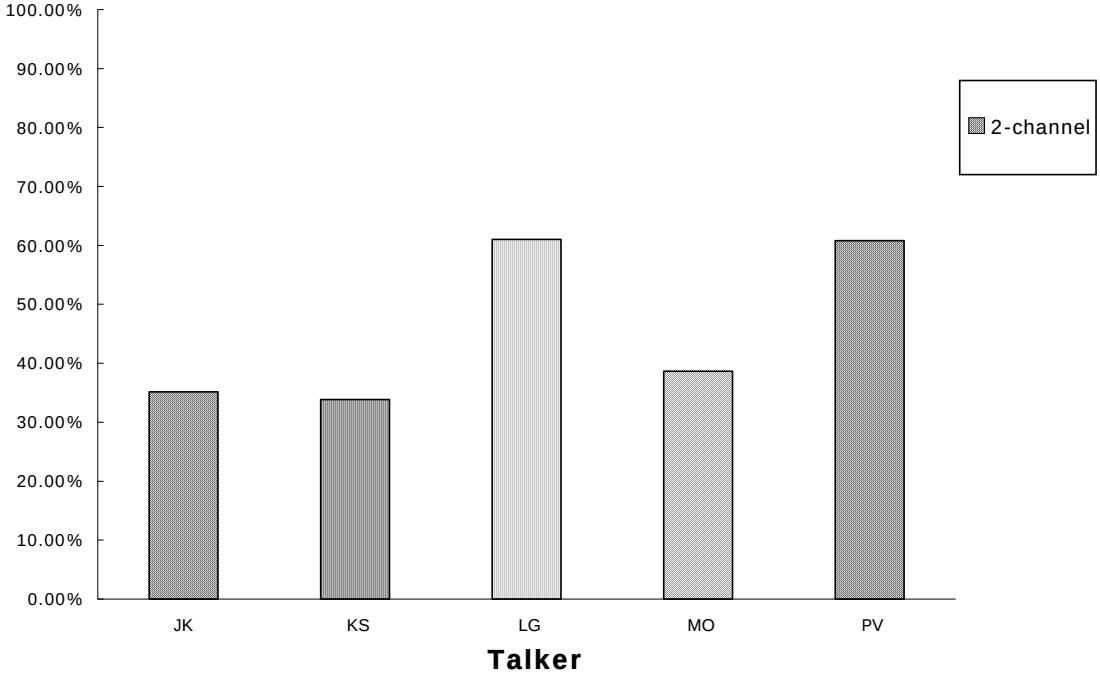
Percent Correct Identification
Figure 1



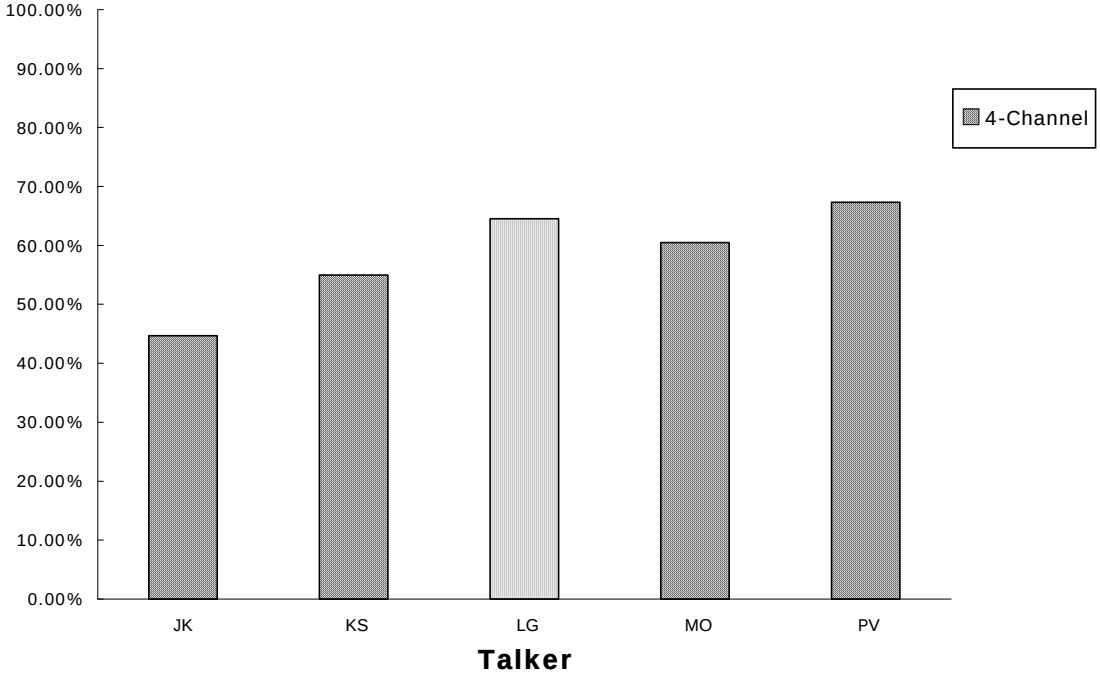
Percent Correct Visual Only by Talker
Figure 2



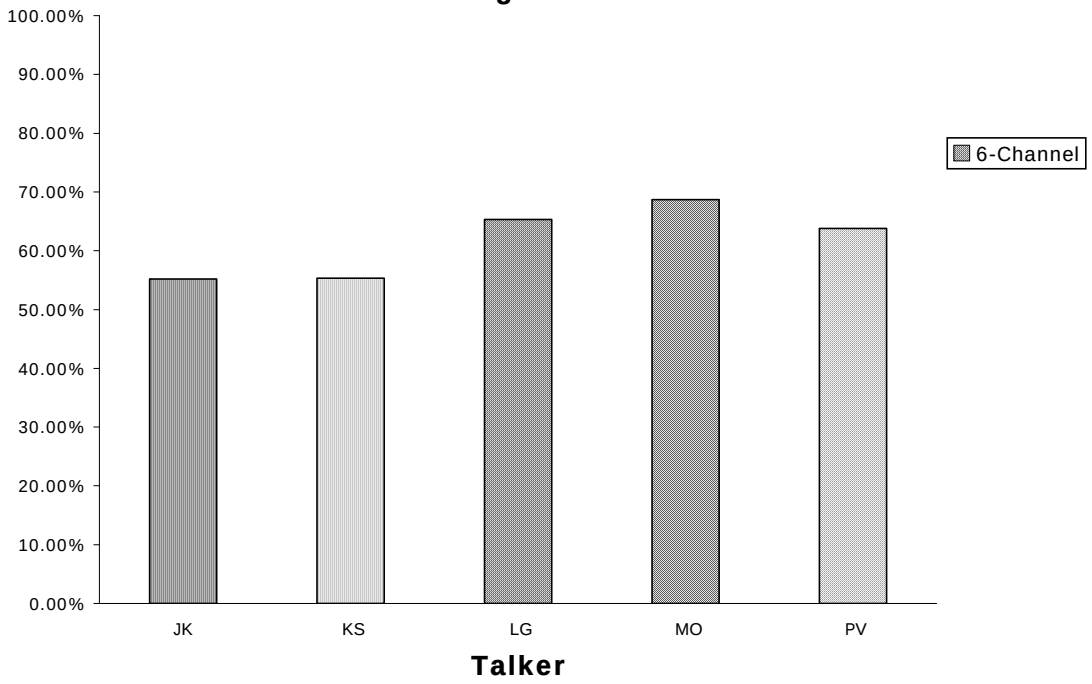
Percent Correct Auditory 2-Channel by Talker
Figure 3



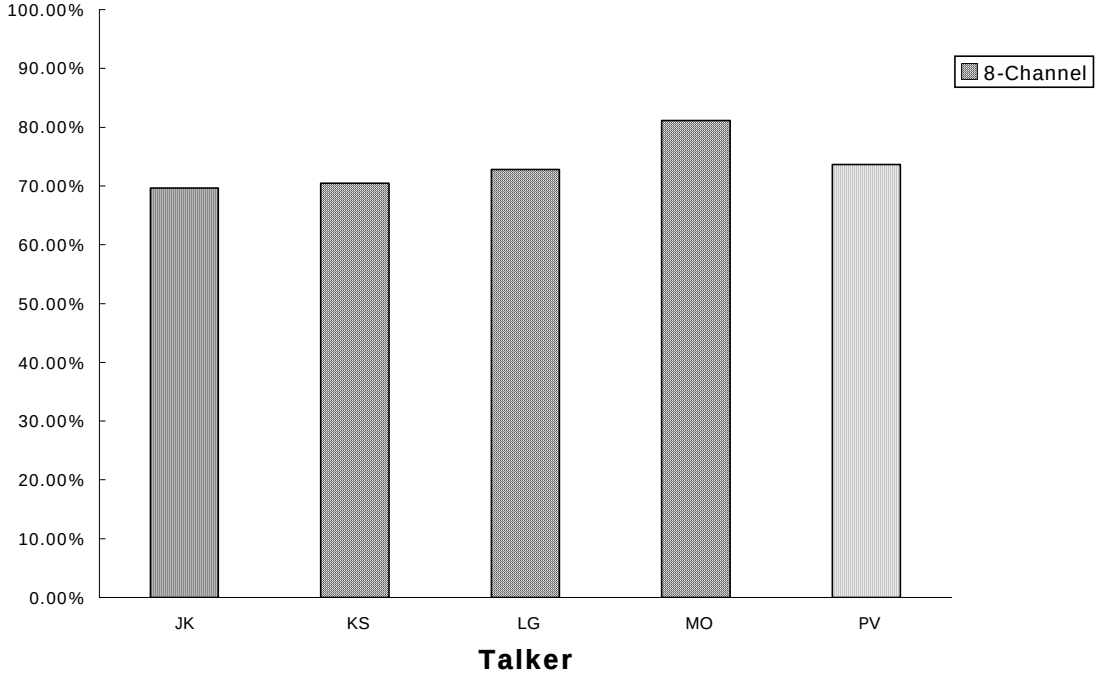
Percent Correct Auditory 4-Channel by Talker
Figure 4



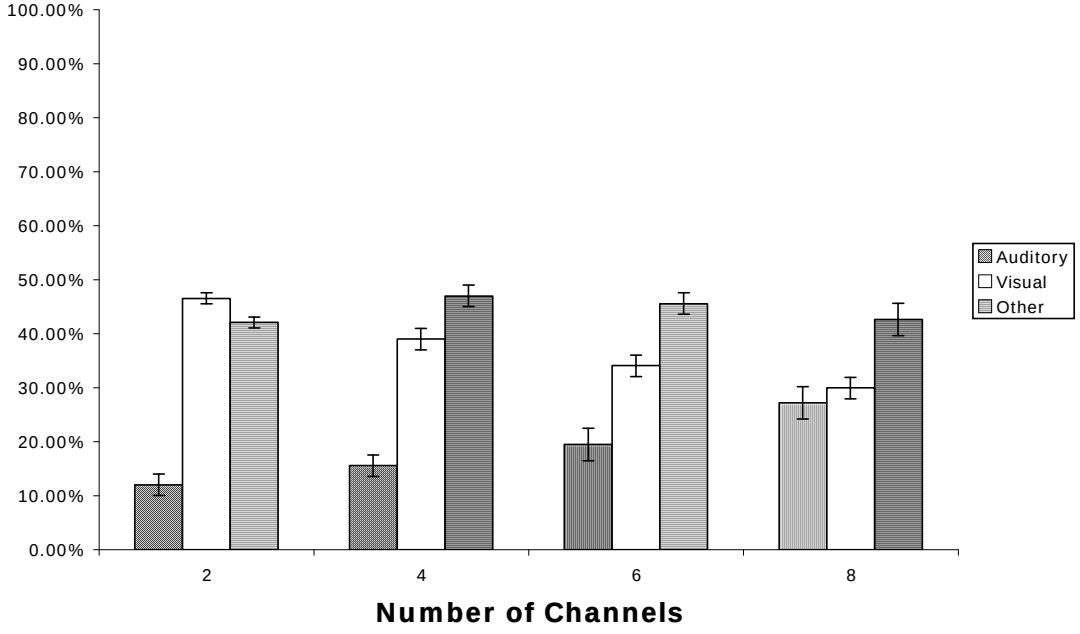
Percent Correct Auditory 6-Channel by Talker
Figure 5



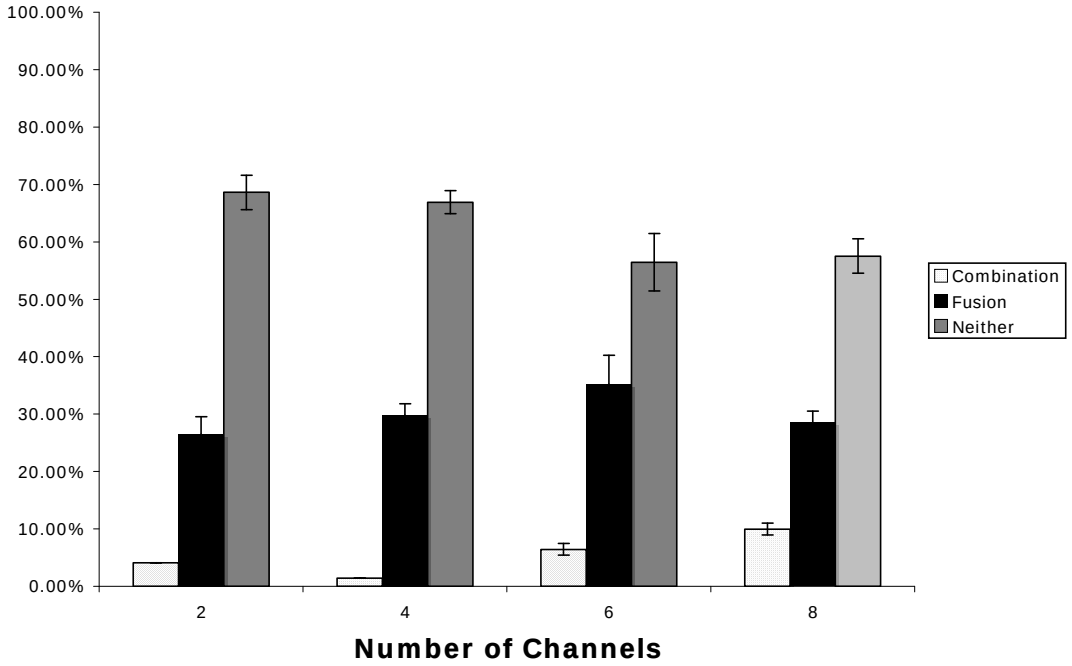
Percent Correct Auditory 8-Channel by Talker
Figure 6



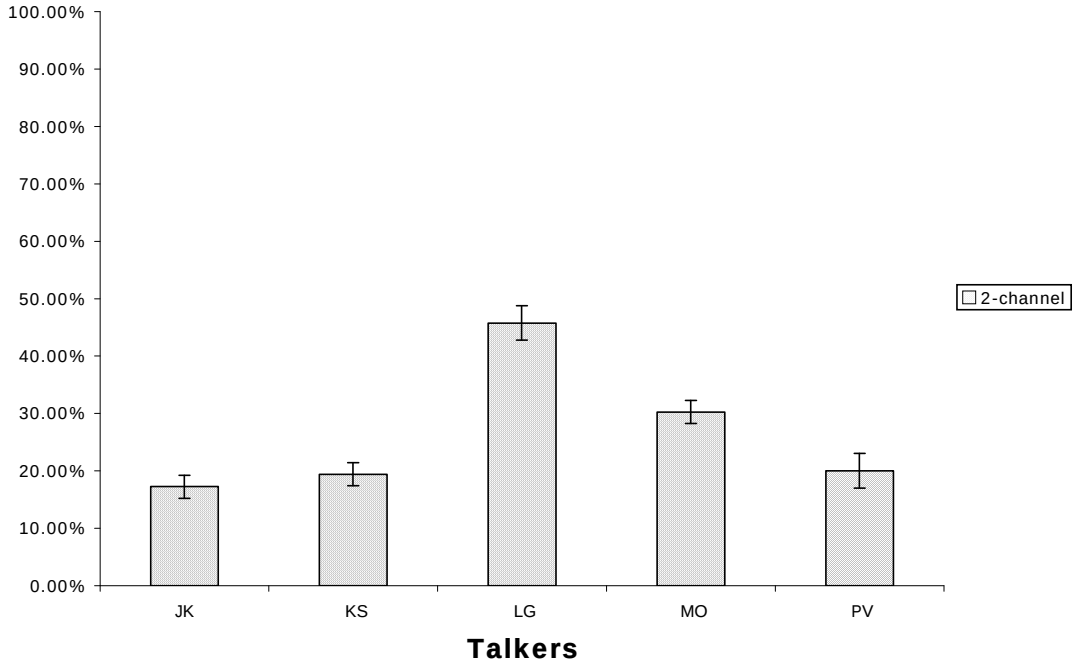
Percent of Auditroy+Visual Integration Responses for Incongruent Inputs
Figure 7



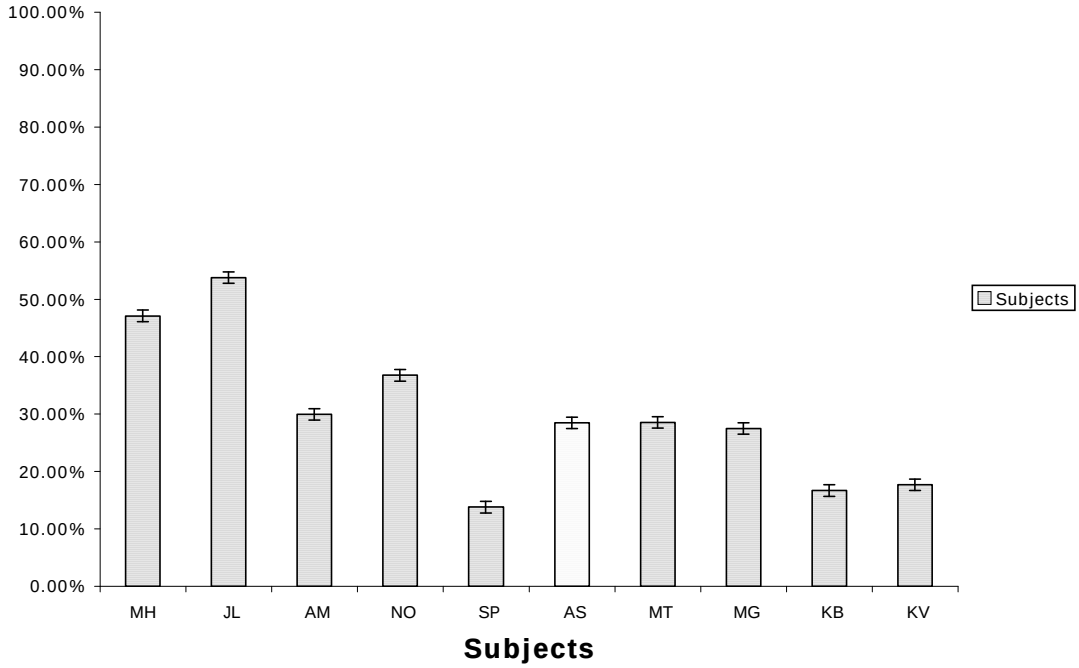
Auditory+Visual Integration for Incongruent Inputs
Figure 8



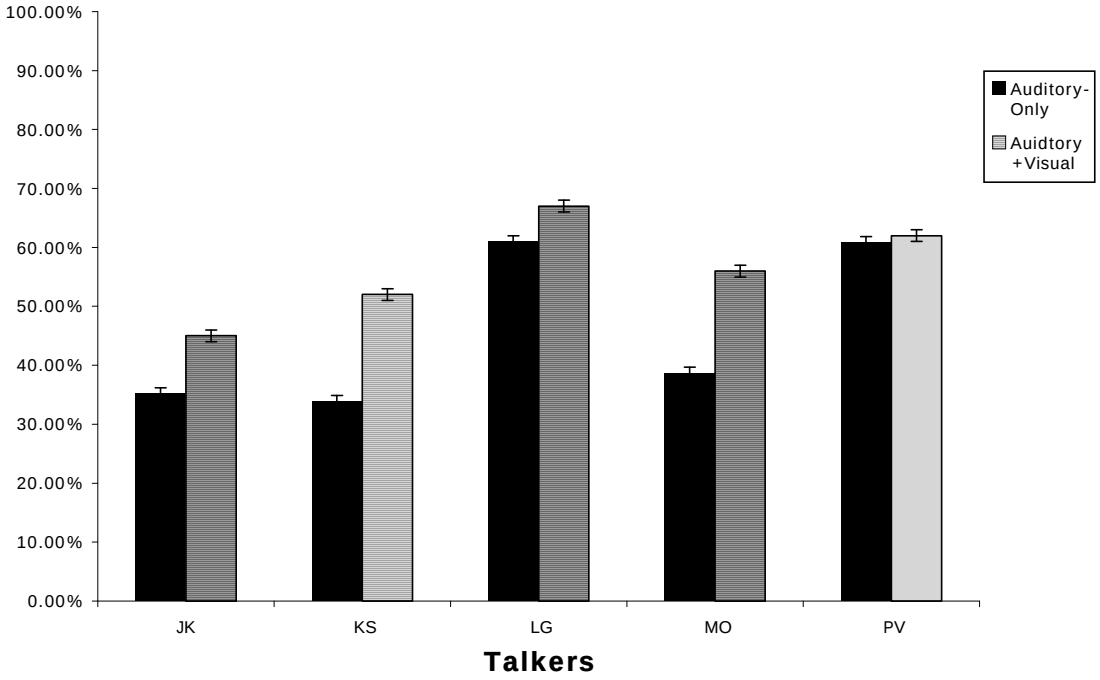
Percent of Fusion Responses 2-Channel by Talker
Figure 9



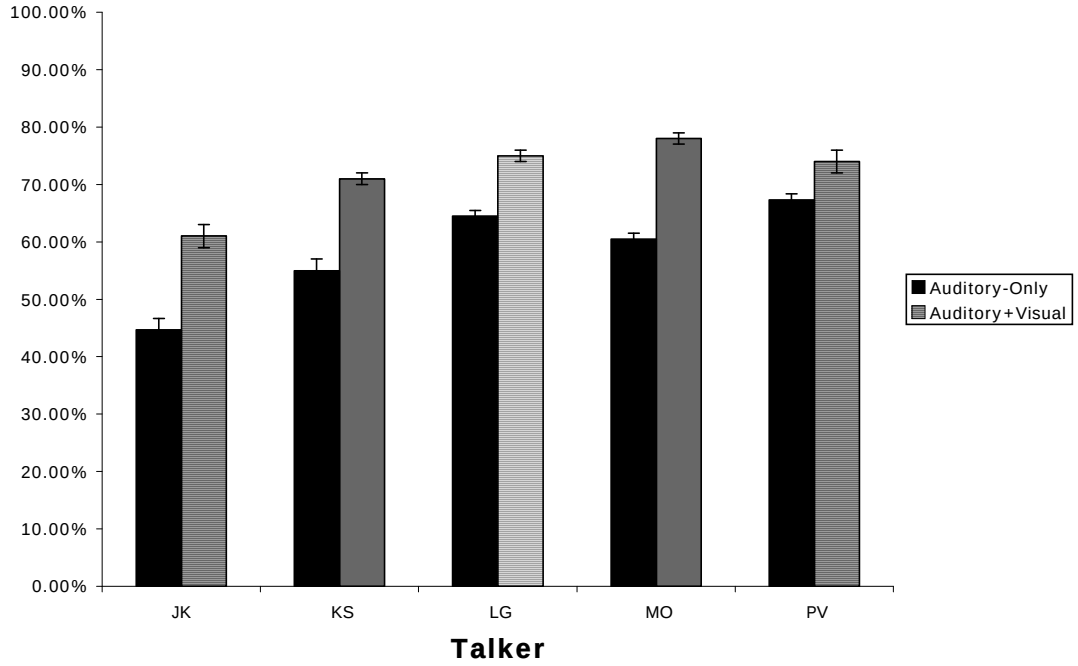
Fusion Responses Across All Channels by Subject
Figure 10



Amount of Integration in the 2-Channel Condition
Figure 11



Amount of Integration in the 4-Channel Condition
Figure 12



Appendix

2 CHANNEL

	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	AT	HAT	FAT	THAT	NAT	DAT	YAT	BGAT	RAT	THETA	BRAT	FLAT	BLAT
BAT	**0.58**	0.02	0.09	0.03	0.02	0.00	0.00	0.01	0.05	0.07	0.06	0.06		0.00		0.00			0.00		
PAT	0.05	**0.55**	0.01	0.01	0.11		0.01	0.01	0.01	0.12	0.08	0.05	0.00				0.00				
MAT	0.05	0.02	**0.73**	0.01	0.02				0.02	0.08		0.05	0.05	0.01			0.01				
GAT	0.30	0.06	0.01	**0.36**	0.10		0.01		0.01	0.06	0.02	0.04		0.02		0.00		0.00	0.01		0.01
CAT	0.02	0.16		0.03	**0.67**		0.04	0.00		0.02	0.04	0.02	0.00								
ZAT	0.13	0.04	0.07	0.03	0.03	**0.08**	0.02	0.02	0.01	0.16	0.14	0.21		0.02	0.03	0.01	0.03	0.01			
TAT	0.02	0.31		0.03	0.32		**0.15**	0.02	0.01	0.04	0.10	0.02		0.01						0.01	
SAT	0.11	0.03	0.01	0.01		0.01	0.01	**0.18**	0.02	0.04	**0.54**	0.06						0.01		0.01	

4 CHANNEL

	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	AT	HAT	FAT	THAT	NAT	DAT	RAT	THETA	LAT	SHAT	DWAT	FLAT
BAT	**0.60**	0.04	0.09		0.01	0.02	0.00		0.03	0.09	0.05	0.06	0.01							
PAT	0.01	**0.69**	0.00		0.02	0.00	0.04	0.00		0.18	0.03	0.01		0.00	0.00					
MAT	0.02	0.02	**0.89**		0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.02							
GAT	0.13	0.05	0.01	**0.46**	0.14	0.00	0.03	0.00	0.01	0.01	0.03	0.04	0.00	0.09			0.00		0.01	
CAT	0.00	0.15	0.01	0.01	**0.68**		0.09			0.03	0.01	0.00							0.01	
ZAT	0.10		0.02		0.01	**0.30**	0.01	0.06		0.02	0.06	0.42	0.01	0.01	0.01	0.01	0.01			
TAT	0.01	0.21			0.15	0.01	**0.56**	0.01		0.02	0.03	0.02				0.01				
SAT	0.08	0.02	0.02		0.02	0.06	0.02	**0.31**		0.02	0.40	0.06				0.02		0.01		0.01

6 CHANNEL

	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	AT	HAT	FAT	THAT	NAT	DAT	RAT	THETA	PCAT	CHAT	VAT
BAT	**0.75**	0.01	0.03	0.00	0.00			0.00	0.03	0.03	0.04	0.09	0.01	0.01					
PAT	0.01	**0.74**	0.01	0.00	0.03		0.02	0.00	0.01	0.14	0.02	0.01		0.00					
MAT	0.03		**0.85**		0.01				0.01	0.03		0.02	0.08						
GAT	0.19	0.01	0.01	**0.46**	0.06		0.01		0.01	0.02	0.03	0.10	0.00	0.11		0.00			
CAT	0.01	0.16		0.01	**0.62**		0.14	0.00	0.00	0.01	0.01	0.02		0.00			0.00	0.00	
ZAT	0.06		0.01	0.01		**0.20**	0.01	0.02		0.01	0.12	0.57			0.02	0.01			0.01
TAT	0.01	0.08			0.08		**0.80**			0.01	0.03	0.01							
SAT	0.04	0.01			0.01	0.01	0.01	**0.48**		0.01	0.39	0.07				0.01			

8 CHANNEL

	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	AT	HAT	FAT	THAT	NAT	DAT	YAT	BGAT	RAT	THETA	LAT	CNAT	WAT	TWAT	LAT
BAT	**0.78**	0.02	0.04		0.00				0.03	0.05	0.02	0.04	0.00	0.00				0.00					
PAT	0.02	**0.78**	0.03		0.02		0.01			0.13	0.00	0.00					0.00						
MAT	0.02	0.01	**0.81**						0.01	0.02		0.03	0.07						0.01		0.03		0.01
GAT	0.09	0.00	0.00	**0.69**	0.05	0.00	0.01	0.00		0.00	0.01	0.02	0.00	0.11		0.00							
CAT		0.11		0.02	**0.76**		0.09			0.01	0.01			0.00						0.00			
ZAT	0.02		0.01	0.01		**0.40**		0.05		0.01	0.08	0.40		0.01	0.01		0.01	0.01			0.02		
TAT		0.05			0.14		**0.78**			0.01	0.01	0.01										0.01	
SAT	0.02		0.01	0.01		0.05		**0.58**			0.25	0.07	0.01					0.04					